МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение
высшего образования
«Уральский федеральный университет имени
первого Президента России Б. Н. Ельцина»

МАКРОСТАТИСТИЧЕСКИЙ АНАЛИЗ И ПРОГНОЗИРОВАНИЕ ДАННЫХ

Лекция № 1

Основные термины и понятия

Екатеринбург
2024

Содержание

Часть 1. Математические модели и имитационное моделирование 3

Часть 2. Понятие временного ряда (time series) 6

Часть 3. Классификация временных рядов.....9

Часть 4. Типовые базовые модели временных рядов 15

Часть 1. Математические модели и имитационное моделирование

Модель – абстрактное описание системы (объекта, процесса, проблемы, понятия) в некоторой форме, отличной от формы их реального существования. **Моделирование** – воспроизведение тех или иных свойств реальных объектов, предметов и явлений с помощью других объектов, процессов, явлений, либо с помощью абстрактного описания. По способу подобного описания тогда выделяют 5 типов моделирования. **Концептуальное моделирование** описывает модель в виде специальных символов и знаков, операций над ними, с помощью синтаксиса описания. **Физическое моделирование** воспроизводит системы исходя из соотношения подобия, вытекающего их схожести тех или иных физических процессов (особенно для периодических, циклических и хаотических явлений). **Структурно-функциональное моделирование** пытается представить модель в виде набора схем, блок-схем, графиков, диаграмм, таблиц, рисунков с описанием их последовательности, то есть в виде строгого набора структурных взаимосвязанных элементов. **Математическое моделирование** – самое известное и строит модели наиболее абстрактно средствами математики и логики, включая средства описания случайных вероятностей в виде их функций распределения. Наконец, **имитационное моделирование** отражает логико-математическую абстрактную модель в виде алгоритма функционирования системы, программно-реализуемой на ЭВМ средствами языка программирования, программных пакетов и т.д.

Несмотря на различные виды моделирования, базой для построения модели все-таки выступает ее математическая схема, которая отражает свойства реальной системы, как некоторый исследуемый «черный ящик», внутренности которого и способ функционирования могут быть, вообще говоря, неизвестны совсем или быть настолько сложны, что не укладываться в выбранный способ моделирования. При наличии множества таких неизвестных элементов, модель объекта математически удобно представлять

как совокупность множества величин, таких как: x – совокупность входных воздействий на систему, v – совокупность воздействий внешней среды, h – совокупность внутренних параметров системы, и y – совокупность выходных характеристик системы. Соотношение между всеми этими множествами устанавливает некоторая функция F, которая описывает либо связи между ними (+ не стоит забывать про параметр времени t), либо представляет собой последовательность состояний моделируемой системы во времени.

По типу всех этих совокупностей, модельному времени и виду функций можно определять самые различные типы математических моделей. Но самая распространенная классификация математических моделей обращает внимание на две ключевые особенности, существенно меняющие способ описания модели. Во-первых, это наличие случайных событий/вероятностей в системе, то есть является ли система **стохастической**, а если случайности нет, то система называется **детерминированной**. Во-вторых, это наличие **непрерывного** или **дискретного** времени (или же уровней/состояний системы), в зависимости от чего у модели будет бесконечное или конечное множество состояний.

Для непрерывно-детерминированного подхода используются математические модели в виде системы дифференциальных уравнений – такие математические модели называют **D**-схемами от слова Differential, по их основной описательной составляющей. В случае же, если система все еще детерминирована (без случайностей), но дискретна (имеет конечное число состояний), то математической моделью служит **F**-схема, от слова Finite Automata, так как такие модели описываются с помощью конечных автоматов. Добавление случайностей в такие автоматы (дискретно-стохастический подход) приводит к **P**-схемам, от слова Probability = вероятность. Самые сложные непрерывно-стохастические системы описываются с помощью **Q**-схем, от слова Queue – ключевого элемента парадигмы систем массового обслуживания (СМО). Ну и кроме подобной упрощенной классификации

существует и целый ряд других моделей, позволяющих описывать разные процессы. Это и **N**-схемы для сетевых процессов, **E**-схемы для оценочных сетей, **A**-схемы для агрегатных, объединяющих, комбинирующих любые другие виды моделей и т.д.

За таким странным понятием как «совокупность воздействий» в математической модели на самом деле скрывается весьма практическое понятие данных. Поток/массив этих данных записывается в реальности со временем и на основе анализа всех доступных данных и строится затем необходимая математическая модель. Поэтому без следующего понятия для решения практических задач нам будет не обойтись.

Часть 2. Понятие временного ряда (time series)

Временной ряд (time series) – упорядоченная последовательность результатов измерений текущих значений одного или нескольких параметров, зафиксированных в последовательные моменты времени. Чаще всего эти показатели являются упорядоченными в хронологическом порядке. Простейшим примером подобного временного ряда может служить ряд курса валют EUR/USD (рис. 1.1).

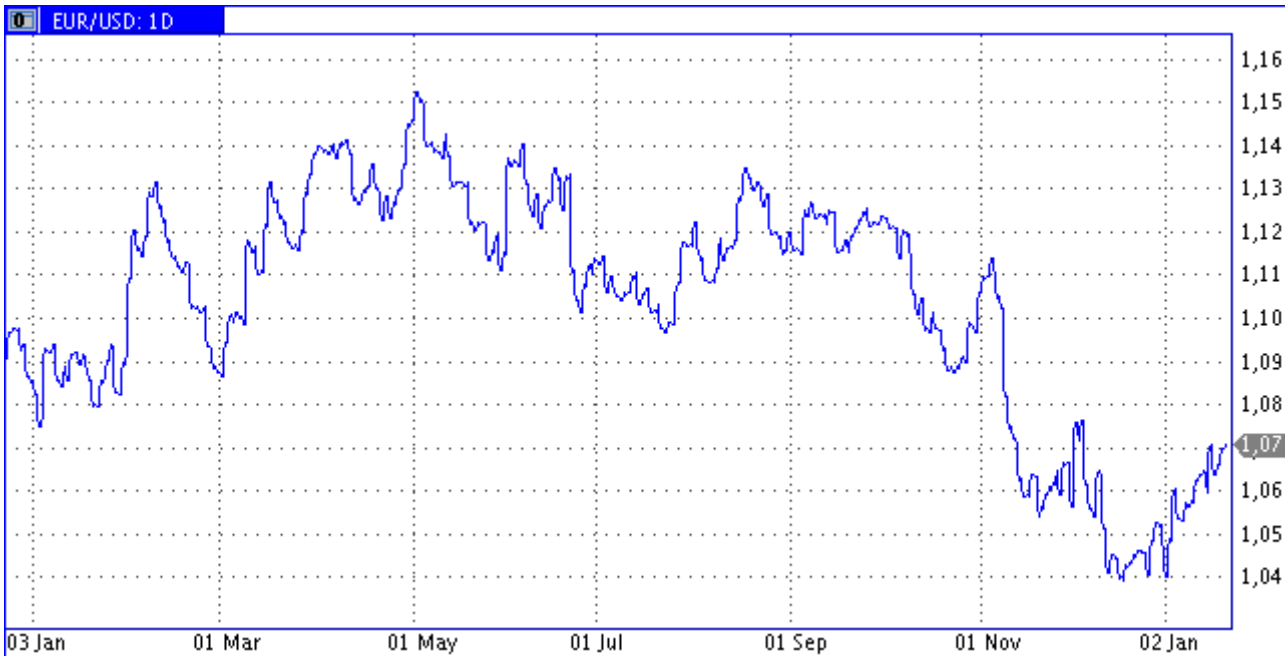

Рисунок 1.1 – Временной ряд, представляющий собой отсчеты показателя курса отношения валют EUR/USD (евро к доллару США)

С точки зрения теории вероятностей временной ряд представляет собой выборку из генеральной совокупности некоторой случайной величины, характеризующейся определенной функцией распределения. Чтобы понять эту фразу, представьте, что есть некоторый процесс, будь то физический или социально-экономический, и он имеет случайный характер и отражается в виде последовательности случайных величин $\xi(t_1),\xi(t_2),...,\xi(t_n)$, где $t_i < t_{i+1}$. Тогда набор наблюдений $\{y_i, y_i = \xi(t_i)\}$ в моменты времени $t_1,t_2,...,t_n$ будет называться **временной выборкой**. В чем же разница между этими понятиями? Дело в том, что этот набор наблюдений уже не является случайным, он уже

зафиксирован и существует в единственном числе. Но по этим наблюдениям в задаче адаптивного анализа данных мы и пытаемся оценить исходный случайный процесс. В дальнейшем под временным рядом (ВР) мы будем понимать именно вот эту временную выборку, но при оценке его параметров в том числе опираться и на лежащую в его основе неизвестную случайную величину.

Изменения случайной величины $\xi(t_i)$ происходят под воздействием самых разнообразных факторов и причин, зачастую скрытых от потенциального наблюдателя. Поэтому в отношении ВР выдвигается предположение, что совокупное влияние факторов формирует некоторую закономерность в эволюции ряда. Поиск этой закономерности различными способами и представляет собой процесс *анализа временных рядов*. Когда подобная закономерность будет найдена, станет возможным построение ряда $\{y_j\}$, который мы еще не наблюдали, но планируем увидеть. Такой процесс построения будущих наблюдений на основе найденной закономерности и проанализированного ВР будет тогда называться процессом *прогнозирования временных рядов*.

При анализе временных рядов можно выделить 4 основных задачи:

1) определение количественных характеристик процесса, породившего данный ВР, который позволят нам понять природу скрытой случайной величины;
2) количественное сравнение ВР друг с другом для выявления сходств и различий между процессами, которыми они порождены;
3) описание формальных моделей, которые могли бы описать эволюцию развития процесса, который породил данный ВР;
4) декомпозиция ВР на элементарные составляющие, которые затем можно анализировать согласно задачам из трех предыдущих пунктов.

Таким образом, следует помнить, что фиксированные наблюдения ВР породил некоторый случайный процесс, но так как они неотрывно связаны

друг с другом, по наблюдениям можно попытаться оценить этот исходный случайный процесс. Как мы увидим в дальнейшем, задача значительно упростится, если часть исходного процесса можно определить как **детерминированную** (то есть такую, что однозначно определяется в зависимости от факторов и текущего момента времени), а часть – как чисто **случайную** (определяемую через собственную функцию распределения). Поиск такого разбиения в настоящее время является самым простым и эффективным способом *анализа* и, одновременно, *прогнозирования* временных рядов. В самом деле, по определению детерминированной составляющей, на ее основе можно будет строить прогноз, а случайная составляющая даст нам все те необходимые статистические характеристики исходного случайного процесса, породившего ВР.

Часть 3. Классификация временных рядов

Всякий временной ряд $\{y_i, y_i = \xi(t_i)\}$, где $t_i < t_{i+1}$, включает в себя два обязательных параметра: это **время** t_i и значение конкретного показателя y_i, называемого **уровнем** или **отсчетом** ряда. По-видимому, от комбинации видов этих двух показателей будет определяться базовая классификация временных рядов. Также еще следует учитывать особенности самого исходного процесса, которые могут существенно ограничить нас в возможностях его анализа. В целом классификация ВР представлена в таблице 1.1 ниже.

Таблица 1.1. Классификация временных рядов

Тип классификации	Виды временных рядов
В зависимости от вида отсчетов	➢ ВР абсолютных величин ➢ ВР относительных величин ➢ ВР средних величин
В зависимости от выбранной временной сетки	➢ Равноотстоящие ряды ➢ Неравноотстоящие ряды
В зависимости от того, как уровни выражают состояния процесса во времени	➢ Интервальные ряды ➢ Моментные ряды
В зависимости от характера случайного процесса, лежащего в его основе	➢ Стационарные ВР ➢ Нестационарные ВР (НВР)

Рассмотрим эти классификации подробнее. Первый тип классификации является самым простым: наблюдения фиксируются либо просто в виде некоторой *абсолютной* величины (в килограммах, метрах, штуках, и т.д.), либо в виде *отношения* двух величин одной размерности (см. рисунок 1.1 EUR/USD), либо наблюдения ведутся постоянно с малым временным интервалом (например, каждый день/час), но фиксируются только в виде *усредненной* величины за более крупный промежуток времени

(среднемесячные/среднегодовые и т.д. значения), так как основной характер процесса определяется крупномасштабными факторами.

Второй вид классификации является более сложным. *Равноотстоящие* ВР формируются при исследовании и фиксации значений процесса в следующие друг за другом **равные** интервалы времени. То есть для заданной временной сетки $\{t_i\}, t_i < t_{i+1}$, значение $t_{i+1} - t_i = \Delta t$ есть **постоянная** величина. Большинство реальных процессов описываются при помощи равноотстоящих временных рядов – просто потому что их проще анализировать, и они имеют более высокую точность методов расчета их характеристик. *Неравноотстоящие* ВР не выполняют принцип равенства интервалов временной сетки, то есть Δt_i меняется. Такие ряды гораздо сложнее анализировать, так как теория анализа подобных ВР намного более слабо отработана на практике, а точность расчета их характеристик снижается. Существует две причины использования подобного неэффективного класса ВР на практике. Во-первых, из-за характера самого исходного процесса: например, все биржевые индексы фиксируют только лишь в рабочие дни недели. Во-вторых, из-за способа фиксации отсчетов ВР: например, в медицине большинство процессов, протекающих в теле человека, возникает спонтанно, и их фиксируют по факту возникновения, привязывая к временной сетке отсчеты уже потом.

Третий вид классификации в чем-то похож на первый. *Интервальный* ВР представляет собой последовательность, в которой абсолютный уровень ряда относят к результату, накопленному или вновь произведенному за определенный интервал времени. Интервальным, например, является ряд показателя выпуска продукции предприятием за неделю, месяц или год; объем электроэнергии, произведенной за час, день, месяц и другие. *Моментные* ВР характеризуются конкретными моментами времени, когда были зафиксированы эти показатели. Примерами моментных рядов являются последовательности финансовых индексов, рыночных цен; физические

показатели, такие как температура окружающего воздуха, влажность, давление, измеренные в конкретные моменты времени, и другие.

Четвертая классификация является самой сложной для понимания, но при этом будет и самой важной. В зависимости от типа ВР по этой классификации будет существенно меняться тип используемых моделей, методика анализа ВР в целом, средства прогноза и т.д. Более того, *адаптивные* средства анализа временных рядов относятся именно к классу нестационарных временных рядов (НВР), так как исторически необходимость в их создании возникла именно из подобного класса задач. Дадим сначала строгие математические определения этих типов ВР, а затем попытаемся понять их с точки зрения практики.

Случайный процесс является **строго стационарным** (или стационарным в **узком** смысле), если его многомерная плотность вероятности

$$p\left(x_1, x_2, ..., x_n, t_1, t_2, ..., t_n\right)$$

произвольной размерности n не изменяется при одновременном сдвиге всех временных сечений $t_1, t_2, ..., t_n$ вдоль оси времени на одинаковую величину τ:

$$p\left(x_1, x_2, ..., x_n, t_1, t_2, ..., t_n\right) = p\left(x_1, x_2, ..., x_n, t_1 + \tau, t_2 + \tau, ..., t_n + \tau\right) \quad (1.1)$$

Если ограничить требования тем, чтобы от временного сдвига не зависели лишь одномерная и двумерная плотности вероятности, то такой случайный процесс называют **стационарным** в **широком** смысле. У стационарного в широком смысле случайного процесса математическое ожидание и дисперсия не зависят от времени, а корреляционная функция зависит только от интервала временной сетки:

$$R_x\left(t_1, t_2\right) = R_x\left(t_2 - t_1\right).$$

Все временные ряды, которые порождены случайными процессами, для которых не выполняется условие стационарности даже в широком смысле, называются **нестационарными временными рядами** (НВР).

Кроме определения стационарности нам еще пригодится подкласс стационарных случайных процессов, называемый **эргодическими** случайными процессами.

Стационарный случайный процесс называется **эргодическим**, если любые из его статистических характеристик, вычисляемых *усреднением по множеству* (ансамблю) реализаций, эквивалентны аналогичным характеристикам, вычисляемым *усреднением по времени* одной, теоретически бесконечно длинной, реализации.

Разберемся, чем же так важны эти два понятия для анализа ВР: стационарность и эргодичность.

Напомним, что любой ряд есть только дискретная выборка из некоторой генеральной совокупности случайной величины. То есть, у нас на руках есть только один экземпляр ВР, который к тому же имеет ограниченную длину по времени, а число отсчетов фиксировано. Большинство статистических, корреляционных и регрессионных методов опирается на различные базовые теоремы и определения из теории вероятностей. В ней статистические характеристики случайных величин всегда *усредняют по ансамблю* (например, тот же метод Монте-Карло основан на этом принципе). С другой стороны, ВР существует только в единственном экземпляре, у нас нет такого множества рядов, чтобы по ним чего-то усреднять. Но, вводя понятие эргодичности, мы можем заменить усреднение по ансамблю на *усреднение по времени*, что избавляет нас от необходимости порождать дополнительные выборки заданного случайного процесса.

От стационарности ВР в большей степени зависят всевозможные статистические критерии и спектральные/корреляционные методы. Связано это с тем, что ключевые алгоритмы, которые строятся в подобных методиках, всегда отталкиваются от исходной оценки мат. ожидания и дисперсии. В той или иной форме обе эти статистические величины участвуют в расчетах и проверке статистических гипотез в виде *постоянных* величин. А следствием

определения стационарности случайного процесса в широком смысле и является как раз тот факт, что мат. ожидание и дисперсия фиксированы и не зависят от времени. Тогда для нестационарных процессов получается, что их мат. ожидание и дисперсия будут являться функциями от времени. Точнее, в зависимости от того отрезка времени ВР, на котором мы оцениваем эти статистические характеристики, будет меняться их результат, то есть характеристики процесса будут зависеть от начала отсчета. Для решения этой проблемы нам либо надо искать такой отрезок времени, на котором мы могли бы считать ряд квазистационарным, либо же использовать методы, которые совсем не требуют предположений о стационарности ВР. Подобные методы анализа НВР как раз называют **адаптивными**: они адаптируют, изменяют свой алгоритм и параметры в зависимости от самого ряда, адаптируют все те величины, что считались постоянными, к некоторым функциям от времени, подобно тому, как мат. ожидание и дисперсия являются функциями от времени для НВР. Отсюда надо понимать, что все результаты, получаемые в ходе анализа НВР адаптивными методами, всегда будут представлены в виде некоторых функций от времени, то есть в виде *мгновенных* величин.

Надо понимать, что на практике большинство ВР оказываются нестационарными, в то время как лучше всего описаны и отработаны стационарные случаи. Стационарные ряды намного проще описывать и анализировать, и, самое главное, прогнозировать, поэтому при анализе любого ВР сначала пытаются оценить, не является ли он стационарным в широком смысле, а затем уже подбирают методику его анализа.

Стоит отметить, что чистое теоретическое определение стационарности случайного процесса (и в узком, и в широком смысле), не применимо на практике. В самом деле, по определению (1.1), чтобы доказать что ВР является стационарным, нам придется перебрать все возможные сдвиги временного интервала (то есть все возможные начала отсчета). На практике используют другой подход: исходный ВР проверяют с помощью определенных

статистических критериев (например, KPSS-test) на гипотезу о стационарности этого ряда. В этом случае ВР будет не абсолютно точно стационарен или нет, а только с некоторой доверительной вероятностью, что позволяет оценить, насколько мы близки к теоретическому широкому смыслу стационарности, насколько верным является решение использовать методы анализа стационарных временных рядов. Понятно, что с этой точки зрения, лучше все ВР изначально считать нестационарными, и использовать для их анализа только адаптивные методы. Но точность этих методов, как мы увидим в дальнейшем, всегда будет проигрывать аналогичным методам, опирающимся на предположение о стационарности случайного процесса, лежащего в основе изучаемого ряда.

Таким образом, еще раз напомним, что вся классификация ВР может быть сведена к следующей схеме (рис. 1.2):

Рисунок 1.2 – Классификация временных рядов

Часть 4. Типовые базовые модели временных рядов

В основе любого ВР лежит некоторый исходный случайный процесс. Его можно описывать несколькими способами: через его функцию распределения и другие статистические характеристики, либо построив для него некоторую математическую вероятностную модель. Первый способ описания хорошо формализуется концептуально, но плохо подходит для описания в виде некоторой программы для ЭВМ. Второй способ гораздо лучше формализуется в виде программы и, потому, более близок для восприятия человеком. По этой причине для анализа ВР реальных процессов различной природы активно применяются различные *модели временных рядов*.

Одним из основных классов моделей временных рядов является класс **аддитивных моделей** вида:

$$y(t_i) = q(t_i) + \xi(t_i), i = 1, 2, ..., N \ , \qquad (1.2)$$

где $y(t_i)$ – отсчеты ВР, $q(t_i)$ – **детерминированная** (неслучайная) составляющая, которая сама по себе может быть представлена некоторой сложной моделью, $\xi(t_i)$ – **случайная** составляющая ВР.

У такой модели есть множества преимуществ, но самое главное из них, это сам факт разделения случайной и детерминированной составляющих. **Детерминированную компоненту** $q(t_i)$ можно использовать для прогноза и описания того процесса, который лежит в основе изучаемого ряда. Варьируя тип этой составляющей можно получить целое семейство различных моделей, подходящих к каждой конкретной задаче.

К **случайной компоненте** $\xi(t_i)$ можно будет отнести все многочисленные факторы случайного характера, влияющие на изучаемый процесс нерегулярно. Эту составляющую еще иногда называют *остаточным рядом*, так как сначала всегда выделяют детерминированную составляющую, а все что в нее не вошло – по определению термина будет случайным. Если изначальной задачей анализа ВР является прогнозирование, то случайную

компоненту $\xi(t_i)$ иногда также называют *шумом*, так как она только препятствует эффективному прогнозу. В любом случае, случайная компонента является обязательной составной частью любого ВР, так как ряд есть только одна выборка из общей генеральной совокупности некоторого случайного процесса, и уникальность этой выборки определяется конкретной реализацией этой компоненты.

Все возможные способы разбиения детерминированной компоненты $q(t_i)$ на части определяют множество подклассов или семейство моделей ВР. В общем виде все эти модели сводятся к следующему выражению:

$$q(t) = w_\tau \cdot \tau(t) + w_s \cdot \sum_j s_j(t) + w_p \cdot \sum_k p_k(t), t \in [0;T] \ , \qquad (1.3)$$

где $\tau(t)$ – **тренд** (или тенденция); $s_j(t)$ – *j*-я **сезонная компонента**, в любом ряду может быть несколько сезонных составляющих (месячные, квартальные, годовые, и т.д.); $p_k(t)$ – *k*-я **периодическая (циклическая) компонента**, которых тоже может быть несколько; w_τ, w_s, w_p – **коэффициенты наличия** или **отсутствия** трендовой, сезонной и периодической составляющих, могут принимать только значения **0** или **1**.

Тренд $\tau(t)$ представляет собой устойчивую закономерность, наблюдаемую в течение длительного интервала времени. Обычно тренд описывается некоторой простой монотонной функцией от времени, либо же содержит в себе в основном все низкочастотные компоненты анализируемого ряда.

Сезонная компонента $s_j(t)$ связана с наличием факторов, действующих с некоторой периодичностью. Это регулярные колебания *квазипериодического* характера, кратные отсчетам временной сетки. Типичными примерами сезонных составляющих являются месячные, квартальные, годовые, и другие повторяющиеся события. Сезонная компонента всегда связана с реальными факторами сезонности из реальной жизни: с праздниками связаны всплески

покупательной способности, время года влияет на сезонность товаров, время созревания продуктов питания определяет их количество и т.д. Сезонных компонент может быть несколько и они, зачастую, кратны между собой, как и в реальной жизни. Важной характеристикой сезонных компонент является тот факт, что они со временем могут *меняться*, то есть имеют **плавающий характер**, поэтому их и называют квазипериодическими. Изменяться у сезонной компоненты могут *амплитудная* составляющая, *фазовая* часть (фазовый сдвиг) или даже сам *период* сезона может плавать возле некоторой средней величины.

Периодическая (циклическая) компонента $p_k(t)$ есть некоторая неслучайная функция, описывающая длительные периоды смены циклов переменной длины и амплитуды. В отличие от сезона цикл имеет фиксированную *тригонометрическую* составляющую, отражающую некоторые циклические изменения в исходном анализируемом процессе.

В настоящее время сезонные и циклические компоненты почти не отличают друг от друга. Во многих современных методиках анализа ВР **сезонными** компонентами называют *квазипериодические* компоненты *малого периода*, с плавающими характеристиками, в то время как *долгосрочные периодические* компоненты называют **циклами**, которые являются гораздо более устойчивыми к изменению своих характеристик. Все, что не входит в эти две тригонометрические составляющие детерминированной модели, как видно из выражения (1.3), обычно принимают за тренд. В этом случае тренд может содержать и долгосрочные периоды/тенденции, но они тогда просто не считаются важными для анализа и прогнозирования данного конкретного процесса.

Наряду с аддитивными моделями (1.2) понятно, что существуют и **мультипликативные** модели:

$$y(t) = q(t) \cdot \xi(t) = \tau(t) \cdot s(t) \cdot p(t) \cdot \xi(t). \qquad (1.4)$$

Кроме того существует и особый класс **комбинированных** моделей, в которых присутствует как аддитивная часть $A(t)$, так и мультипликативная часть $M(t)$, причем комбинации могут быть самыми разными:

$$y(t) = M(t) + A(t) = \tau(t) \circ s(t) \circ p(t) + \xi(t). \quad (1.5)$$

Другим важным подмножеством моделей, которые будут рассматриваться более подробно позднее, являются так называемые **авторегрессионные** модели:

$$y(t_i) = f\left(y(t_{i-1}), y(t_{i-2}), \ldots\right) + g\left(a_i, a_{i-1}, a_{i-2}, \ldots\right) + \xi(t_i)\ . \quad (1.6)$$

где функция f отражает характер взаимосвязи между последующими и предыдущими значениями ВР, а функция g отражает связь последующих значений ВР от предыдущих шумов. Для широкого круга процессов обе эти функции f и g имеют линейный характер, тогда такие модели называют *смешанной линейной авторегрессией*:

$$y(t_i) = \phi_1 y(t_{i-1}) + \phi_2 y(t_{i-2}) + \ldots + \phi_p y(t_{i-p}) - a_i - \theta_1 a_{i-1} - \ldots - \theta_q a_{i-q} + \xi(t_i). \quad (1.7)$$

В остальном, все эти модели имеют похожую структуру.